CORE
Metadata, citation and similar papers at core.ac.uk
Provided by International Journal of Advances in Intelligent Informatics

International Journal of Advances in Intelligent Informatics
Vol. 4, No. 1, March 2018, pp. 70-75
ISSN 2442-6571

Variable precision rough set model for attribute selection on environment impact dataset

Ani Apriani [a,1,*], Iwan Tri Riyadi Yanto [b,2], Septiana Fathurrohmah [c,3], Sri Haryatmi [d,4], Danardono [d,5]

[a] Department of Geology Engineering, STTNAS, Yogyakarta, Indonesia
[b] Department of Information System, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
[c] Department of Urban and Regional Planning, STTNAS, Yogyakarta, Indonesia
[d] Department of Mathematics, Universitas Gajah Mada, Yogyakarta, Indonesia
[1] aniapriani@sttnas.ac.id; [2] yanto.itr@is.uad.ac.id; [3] septiana@sttnas.ac.id; [4] s.kartiko@yahoo.com; [5] danardono@ugm.ac.id
* corresponding author

ARTICLE INFO

Article history
Received September 9, 2017
Revised March 13, 2018
Accepted March 31, 2018

Keywords
Environment
VPRS
Error classification
Attribute selection

ABSTRACT

The investigation of environment impact have important role to development of a city. The application of the artificial intelligence in form of computational models can be used to analyze the data. One of them is rough set theory. The utilization of data clustering method, which is a part of rough set theory, could provide a meaningful contribution on the decision making process. The application of this method could come in term of selecting the attribute of environment impact. This paper examine the application of variable precision rough set model for selecting attribute of environment impact. This mean of minimum error classification based approach is applied to a survey dataset by utilizing variable precision of attributes. This paper demonstrates the utilization of variable precision rough set model to select the most important impact of regional development. Based on the experiment, The availability of public open space, social organization and culture, migration and rate of employment are selected as a dominant attributes. It can be contributed on the policy design process, in term of formulating a proper intervention for enhancing the quality of social environment.

This is an open access article under the CC–BY-SA license.

1. Introduction

Development is a process of social changes. It aims to enhance the society livelihood, without jeopardizing its environmental and cultural sustainability. Development may enable people within a society to decide their own future. In other words, it has to be participative. Based on those standpoints of views, development should elaborate all of the economic, social, and environmental aspects. Those aspects are inevitably important within the concept of sustainable development. Nevertheless, those three aspects are also the most vulnerable part which could be influenced by the side-effect of development. The imbalance development process may cause negative effect on the either economic, social, or environmental aspects of a society[1]–[3].

The development is often assumed as the physical development in a certain area. It is often marked by the infrastructure and facility enhancement in order to fulfill the social and economic needs of a society. One of the obvious sign of physical development is the land conversion: either convert unused land into building and structure or land function conversion, such as from residential into commercial use. Physical development may cause economic, social, and environmental shifts[4], [5].

The physical development may raise both positive and negative impact in society. The increasing commercial activities, for instance, could provide financial benefits for the society[6]–[9]. However, it may also lead into a higher social-economic disparity within a society. Therefore, this research aims to investigate the social, economic, and environmental impact of the physical development. A strategic program can be well planned by evaluating the environmental impact during the study period in an institution[10]–[12].

An effective way to detect the most principal impact is the use of data mining technique[13]. Data mining in general, is the process of finding, analyzing a new information that may exist in data and summarizing the results as useful information. There are many outstanding studies on data mining in the many areas, such as clustering, association rules, classification, and conflict analysis [14]–[16].

In order to achieve the research objective, this research presents the utilization of the variable precision from rough set (VPRS) theory to perform attribute selection in environment dataset. This method was based on variable precision rough set approximation using minimum error classification of attributes [17]. The VPRS is introduced by Ziarko [18]. The extension of rough set may solve uncertainty data without any functional relationship attributes using error-tolerance capability [19]. By setting the tolerance, VPRS also can handle the noisy data. This paper contribution is the selection of the most influential attribute by ordering the attributes based on its relevancies in term of the attribute minimum error classification of attribute in the dataset. Selecting and identifying the most influential attribute of the environment data set in the early phase could contribute to a better policy design process for the purpose of the enhancement of the social environment.

This paper consists of four sections. The first section is the introduction. It is followed by the method which is include the theoretical review on the concept of information system, rough set data theory, and minimum error classification in section 2. The description of the data set characteristic will be presented in section 3, is also discussed the result of the experiment and the evaluation of the experiment will be discussed. The last section, the conclusion of the paper is in section 4.

2. Method

2.1. Set approximations of information system

The set of approximations of information system can be defined in the following terms:

Definition 1. Let U is Universe, A is attribute set and $V = \bigcup_{\{a \in A\}} V_a$ is the domain of attribute a. An information system can be defined as a mapping of pair universe U and attribute A to value set V [20][19], as in (1).

$$f: U \times A \to V \tag{1}$$

such that $f(u, a) \in V_a, \forall (u, a) \in U \times A$.

Definition 2. Two elements $x, y \in U$ are said to be B-indiscernible $\Leftrightarrow f(x, a) = f(y, a), \forall a \in B$.

A unique indiscernibility relation can be induced for every subset of$A$. Let $B \subset A$, $IND(B)$ is a relation of indiscernibility induced by the set of attribute B and it is an equivalence relation. The partition of U induced by $IND(B)$ is denoted by U/B and $[x]_B$ is the equivalence class in the partition U/B containing $x \in U$.

Definition 3. Lower and upper approximation of X induced by B are defined as in (2).

$$\underline{B}(X) = \{x \in U: [x]_B \subseteq X\} \text{ and } \overline{B}(X) = \{x \in U: [x]_B \cap U \neq \emptyset\} \tag{2}$$

The accuracy of approximation of any subset $X \subseteq U$ with respect to $B \subseteq A$ is defined as in (3).

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|} \tag{3}$$

2.2. Variable Precision Rough Set (VPRS)

Variable precision rough set is extension of rough set theory. This is established by relaxing the subset operator. It is utilized to conduct analysis and identification of patterns of data that represented statistical trends rather than functional. VPRS classifies object based on its smaller error compare with the certain pre-defined level. The threshold introduced in this method does not require any information besides which is already in the data. In VPRS firstly introduced the error classification to define the lower and upper approximation[18]. The error classification in VPRS is defined as in (4).

$$e(X,Y) = \begin{cases} 1 - \frac{|X \cap Y|}{|X|} & ,|X| > 0 \\ 0 & ,|X| = 0 \end{cases}, \quad (4)$$

for every $X, Y \subseteq U$, where $X, Y \neq \emptyset, e(X,Y)$ is called the error classification rate of X relative to Y.

Definition 4. Lets U is a universe and $X \subseteq U$. A real number $0 \leq \beta \leq 0.5$ is given as a threshold. The lower and upper approximation of X are defined as in (5) and (6), respectively.

$$\underline{B}_\beta(X) = \{x \in U : e([x]_\beta, X) \leq \beta\} \quad (5)$$

$$\overline{B}_\beta(X) = \{x \in U : e(|x|_\beta, X) < 1 - \beta\} \quad (6)$$

The Equation (5) is also called as the positive region of X that is the set object of U which can be classified into X with error classification rate not greater than β. The we have $\underline{B}_\beta(X) \subseteq \overline{B}_\beta(X) \Leftrightarrow 0 \leq \beta \leq 0.5$, where β is restricted in interval [0,0.5] to keep the meaning of upper and lower approximation. The accuracy of VPRS with given threshold β is presented in (7).

$$\alpha_{B_\beta}(X) = \frac{|\underline{B}_\beta(X)|}{|\overline{B}_\beta(X)|} \quad (7)$$

2.3. Minimum Error Classification (MECC)

Minimum error classification is a technique that is established by approximating the attribute roughness in the VPRS. This approximation is set by introducing the threshold β which respect to the error classification. By introducing the threshold, β, the accuracy of approximation for selecting attribute clustering increase as it is shown by using definition 4. It can bee see if $\beta > 0.5$ then $\underline{B}_\beta(X) \not\subset \overline{B}_\beta(X)$. However, if $0 \leq \beta \leq 0.5$, then $\underline{B}_0(X) \supseteq \underline{B}_\beta(X)$ and $\overline{B}_0(X) \subseteq \overline{B}_\beta(X)$. Thus, $|\underline{B}_0(X)| \leq |\underline{B}_\beta(X)|$ and $|\overline{B}_0(X)| \geq |\overline{B}_\beta(X)|$. Obviously for $\beta = 0$, the accuracy $\alpha_B(X) = \alpha_{B_\beta}(X)$. On the other hand, for $0 < \beta < 0.5$, $\alpha_B(X) \leq \alpha_{B_\beta}(X)$.

Let $S = (U, A, V, f)$ be an information system. Suppose that $a_i \in A, V(a_i)$, has k-different values, say $y_k, k = 1,2,\wedge,n$. Let $X(a_i = y_k)$, $k = 1,2,\wedge,n$ is a subset of the objects that having k-different values of attribute a_i. The error classification rate of $(a_i = y_k)$ relative to$(a_j = y_k)$,where $i \neq j$, can be defined as in (8).

$$e\left(X(a_i = y_k), X(a_j = y_k)\right) = 1 - \frac{\left|X(a_i = y_k) \text{I} \; X(a_j = y_k)\right|}{\left|X(a_i = y_k)\right|}. \quad (8)$$

The problem appears is how to choose the threshold β to increase the approximation accuracy while the error classification could be minimized. There are three cases of B derived from proposition (8):

Case 1. If $\beta \geq 0.5$, it is clear the accuracy will be out.

Case 2. If $\beta = 0$, then $\alpha_B(X) = \alpha_{B_\beta}(X)$, the accuracy is equal.

Case 3. If $0 < \beta < 0.5$, Hence $\alpha_B(X) < \alpha_{B_\beta}(X)$, the accuracy of VPRS is greater than the traditional rough set.

Based on the previously mentioned cases, β can be formulated as positive number with the value is less than 0.5. The threshold $\beta > 0$ can be selected as the minimum error classification which is denoted as (9).

$$\beta = \arg\min\left\lfloor mean\left\{e\left(X\left(a_i = y_k\right), X\left(a_j = y_k\right)\right)\right\}\right\rfloor \quad (9)$$

The attribute with minimum $\beta > 0$ is selected as the clustering decision. The pseudo code of the MECC algorithm is shown in Fig. 1.

```
Algorithm: MECC
Input: Data
Output: Selecting attribute
Begin
    Step 1. Compute the equivalence classes using the indiscernibility
            relation on each attribute.
    Step 2. Determine the error classification of attribute a1 with
            respect to all aj, where i ≠ j.
    Step 3. Select the mean error classification from step 2 to be a β.
    Step 4. Select an attribute based on the minimum of the β
End
```

Fig. 1. The MECC algorithm

3. Results and Discussion

This research aims to identify the most influential impact from the environment dataset by put the relevant attributes in a ranked order based on the minimum error classification in the dataset. The selection and identification of the most influential attribute of environmental impact in the early stage could help the policy maker to design the proper intervention and take immediate action to improve the quality of social environment.

The dataset was established by conducting a survey in Yogyakarta, Indonesia. There are 400 respondents who involved in the survey. The respondent consists of 176 male and 224 female respondents. Reliability test has been conducted for the data set, with alpha score of 0.953. Data collected from survey is accumulated. The descriptive statistics, which is calculated by utilizing SPSS software, of the data set in form of mean and standard deviation is presented in Table 1. The impact of Physics and chemical, Socio-culture, economic are independent one another, thus the means of its impacts are deferent that are 25.505, 12.7175 and 16.7475, respectively. Meanwhile, the dispersion of the impact is practically homogent, where the standard deviations are quite similar that are 2.954, 1.579 and 2.667, respectively.

Table 1. Mean and standard deviation of variables

	Physic and chemical	Socio-culture	economic
Mean	25.505	12.7175	16.7475
Standard deviation	2.954963	1.579218	2.667683

The average of classification error values for each attributes of the three aspects, can be described in the following part:

1) Impact on Physic and chemical Aspects

There are nine attributes of the physics and chemical aspects, namely: water quantity (L1), water quality (L2), water absorption level (L3), temperature (L4), air pollution level (L5), climate (L6), noise

level (L7), land use (L8), availability of public open space (L9). The average of minimum error classification values are shown in Table 2. The selected attribute, availability of public open space (L9), has the minimum error classification is 0.5.

Table 2. The average of classification error value of each attribute of environmental aspect

Attribute	L1	L2	L3	L4	L5	L6	L7	L8	L9
MECC	0.8	0.6	0.8	0.6	0.6	0.6	0.6	0.8	0.5

2) Impact on Socio-Cultural Aspects

There are five attributes of the socio-cultural aspects, namely: social organization (SB1), social interaction (SB2), culture (SB3), social practice (SB4), livelihood quality (SB5). The average of minimum error classification values can be seen in Table 3. The selected attribute is social organization (SB1) and culture (SB3) with the minimum error classification is 0.6, respectively.

Table 3. The average of classification error value of each attribute of socio-cultural aspects

Attribute	SB1	SB2	SB3	SB4	SB5
MECC	0.6	0.7	0.6	0.7	0.7

3) Impact on economic aspects

There are ten attributes of the "economic aspects", namely: migration (E1), rate of employment (E2), economic structure development (E3), revenue (E4), expenditure (E5), shift of occupation (E6), public health (E7), increasing number of educational facility (E8), increasing number of religious facility (E9), increasing number of health care facility (E10). The average of minimum error classification values are shown in Table 4. The selected attributes are migration (E1), and rate of employment (E2) which both have the minimum error classification is 0.49.

Table 4. The average of classification error value of each attribute of economic aspect

Attribute	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10
MECC	0.49	0.49	0.75	0.75	0.75	0.5	0.75	0.5	0.5	0.5

4. Conclusion

This paper demonstrates the utilization of the variable precision rough set as attribute selection to the environmental impact. The utilization of the mean of minimum error classification using variable precision of attributes is the basic of this technique. This technique was utilized to examine three environment impacts, such as Physical and Chemical aspects, socio-cultural aspects, and economic aspects. This paper has demonstrated the usefulness of this technique to select the most influential environment impact. The selected attributes are availability of public open space, social organization and culture, migration and rate of employment. The result from the experiment as it has been presented in this paper can be a basic for the policy design process and to formulate the proper treatment to improve the quality of social environment.

Acknowledgment

This work was supported by Directorate of Research and Community Service Kemenristekdikti with grant number is 58/STTNAS/P3M/Pen.Dikti/V/2017.

References

[1] M. T. Dugan, E. H. Turner, M. A. Thompson, and S. M. Murray, "Measuring the financial impact of environmental regulations on the trucking industry," *Res. Account. Regul.*, vol. 29, no. 2, pp. 152–158, 2017, doi: https://doi.org/10.1016/j.racreg.2017.09.007.

[2] P. Ashcroft and L. Murphy Smith, "Impact of environmental regulation on financial reporting of pollution activity: A comparative study of U.S. and Canadian firms," *Res. Account. Regul.*, vol. 20, pp. 127–153, 2008, doi: https://doi.org/10.1016/S1052-0457(07)00207-X.

[3] C. Mary Schooling, E. W. L. Lau, K. Y. K. Tin, and G. M. Leung, "Social disparities and cause-specific mortality during economic development," *Soc. Sci. Med.*, vol. 70, no. 10, pp. 1550–1557, 2010, doi: https://doi.org/10.1016/j.socscimed.2010.01.015.

[4] J. K. Woo, D. S. H. Moon, and J. S. L. Lam, "The impact of environmental policy on ports and the associated economic opportunities," *Transp. Res. Part A Policy Pract.*, no. xxxx, pp. 0–1, 2017, doi: https://doi.org/10.1016/j.tra.2017.09.001.

[5] L. M. Ferri and M. Pedrini, "Socially and environmentally responsible purchasing: Comparing the impacts on buying firm's financial performance, competitiveness and risk," *J. Clean. Prod.*, vol. 174, pp. 880–888, 2018, doi: https://doi.org/10.1016/j.jclepro.2017.11.035.

[6] J. K. Owusu-Ansah and F. Atta-Boateng, "The spatial expression of physical development controls in a fast growing Ghanaian city," *Land use policy*, vol. 54, pp. 147–157, 2016, doi: https://doi.org/10.1016/j.landusepol.2016.02.012.

[7] A. Kumari and A. K. Sharma, "Physical & social infrastructure in India & its relationship with economic development," *World Dev. Perspect.*, vol. 5, pp. 30–33, 2017, doi: https://doi.org/10.1016/j.wdp.2017.02.005.

[8] P. Clavel and R. Young, "'Civics': Patrick Geddes's theory of city development," *Landsc. Urban Plan.*, vol. 166, no. June, pp. 37–42, 2017, doi: https://doi.org/10.1016/j.landurbplan.2017.06.017.

[9] S. Ullrich-French, A. N. Cole, and A. K. Montgomery, "Evaluation development for a physical activity positive youth development program for girls," *Eval. Program Plann.*, vol. 55, pp. 67–76, 2016, doi: https://doi.org/10.1016/j.evalprogplan.2015.12.002.

[10] A. K. M. Tarigan, D. A. A. Samsura, S. Sagala, and A. V. M. Pencawan, "Medan City: Development and governance under the decentralisation era," *Cities*, vol. 71, no. July, pp. 135–146, 2017, doi: https://doi.org/10.1016/j.cities.2017.07.002.

[11] A. K. M. Tarigan, D. A. A. Samsura, S. Sagala, and R. Wimbardana, "Balikpapan: Urban planning and development in anticipation of the post-oil industry era," *Cities*, vol. 60, pp. 246–259, 2017, doi: https://doi.org/10.1016/j.cities.2016.09.012.

[12] W. Li, C. Wu, and S. Zang, "Modeling urban land use conversion of Daqing City, China: a comparative analysis of 'top-down' and 'bottom-up' approaches," *Stoch. Environ. Res. Risk Assess.*, vol. 28, no. 4, pp. 817–828, May 2014, doi: https://doi.org/10.1007/s00477-012-0671-0.

[13] T. Woldai and A. G. G. Fabbri, "The Impact of Mining on The Environment," in *Deposit and Geoenvironmental Models for Resource Exploitation and Environmental Security*, 2002, pp. 345–364, doi: https://doi.org/10.1007/978-94-010-0303-2_17.

[14] Hamdani and R. Wardoyo, "A Review on fuzzy multi-criteria decision making land clearing for oil palm plantation," *Int. J. Adv. Intell. Informatics*, vol. 1, no. 2, pp. 75–83, 2015, https://doi.org/10.26555/ijain.v1i2.26.

[15] M. Muhajir and B. R. Efanna, "Association Rule Algorithm Sequential Pattern Discovery using Equivalent Classes (SPADE) to Analyze the Genesis Pattern of Landslides in Indonesia," *Int. J. Adv. Intell. Informatics*, vol. 1, no. 3, pp. 158–164, 2015, doi: https://doi.org/10.26555/ijain.v1i3.50.

[16] D. Ismi, S. Panchoo, and M. Murinto, "K-means clustering based filter feature selection on high dimensional data," *Int. J. Adv. Intell. Informatics*, vol. 2, no. 1, pp. 38–45, 2016, doi: http://dx.doi.org/10.26555/ijain.v2i1.54.

[17] I. T. R. Yanto, R. R. Saedudin, D. Hartama, and T. Herawan, *Clustering based on classification quality (CCQ)*, 2017, vol. 549 AISC, doi: http://dx.doi.org/10.1007/978-3-319-51281-5_33.

[18] W. Ziarko, "Variable precision rough set model," *J. Comput. Syst. Sci.*, vol. 46, no. 1, pp. 39–59, 1993, doi: https://doi.org/10.1016/0022-0000(93)90048-2.

[19] Z. Pawlak, "Rough sets," *Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, 1982, doi: http://dx.doi.org/10.1007/BF01001956.

[20] Z. Pawlak and A. Skowron, "Rudiments of rough sets," *Inf. Sci. (Ny).*, vol. 177, no. 1, pp. 3–27, Jan. 2007, doi: http://dx.doi.org/10.1016/j.ins.2006.06.003.